Composante du produit nº 82-003-X au catalogue de Statistique Canada
Rapports sur la santé

Article

Combiner les cycles de l'Enquête sur la santé dans les collectivités canadiennes

Par Steven Thomas et Brenda Wannell

Février 2009

Statistique Canada Statistics Canada

Canada

Combiner les cycles de l'Enquête sur la santé dans les collectivités canadiennes

par Steven Thomas et Brenda Wannell

Résumé

Contexte
Un seul cycle de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) ne répond pas toujours aux besoins analytiques des chercheurs. Le présent article décrit des méthodes de combinaison des divers cycles de l'ESCC et discute des problèmes dont il convient de tenir compte si ces données sont combinées. Un exemple empirique illustre les méthodes proposées.

Données et méthodes
Deux méthodes peuvent être utilisées pour combiner les cycles de l'ESCC : la méthode individuelle et la méthode groupée. Dans le cas de la méthode individuelle, les estimations sont calculées pour chaque cycle séparément, puis combinées. Dans le cas de la méthode groupée, les microdonnées sont combinées et l'ensemble de données résultant est traité comme s'il provenait d'un échantillon d'une seule population.

Résultats
Dans le cas de l'approche individuelle, il est recommandé d'utiliser la simple moyenne des estimations. Pour l'approche groupée, il est conseillé de rééchelonner les poids en les multipliant par un facteur constant afin de pouvoir créer une estimation de période couvrant les périodes couvertes par les cycles individuels. Le choix de la méthode dépend du but de l'analyse et des données disponibles.

Interprétation
La combinaison des cycles ne devrait être envisagée que si les estimations pour la période la plus récente ne suffisent pas. Les deux méthodes obscurcissent les tendances d'un cycle à l'autre et ne révèlent pas les changements de comportement liés aux initiatives en matière de santé publique.

Mots-clés

Collecte des données, regroupement des données, méta-analyse, interprétation des données statistiques.

Auteurs

Steven Thomas (1-613-951-7300; Steven.Thomas@statcan.gc.ca) travaille à la Division des méthodes d'enquête auprès des ménages et Brenda Wannell (1-613-951-8554; Brenda.Wannell@statcan.gc.ca) travaille à la Division de la statistique de la santé à Statistique Canada, Ottawa (Ontario) K1A 0T6.

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) comprend deux enquêtes par sondage transversales. Le cycle .1 a pour but de recueillir des renseignements généraux sur la santé dans plus de 120 régions sociosanitaires, tandis que le cycle .2 porte sur des aspects particuliers de la santé et a pour but de recueillir des données pour la production d'estimations au niveau provincial.

Malgré les grandes tailles d'échantillon, un seul cycle de l'ESCC ne permet pas nécessairement de répondre aux besoins des utilisateurs. Par exemple, certains chercheurs pourraient vouloir étudier une population rare, définie par des données géographiques détaillées ou par des caractéristiques sociodémographiques ou des caractéristiques de la santé rares. Comme un cycle unique ne fournit parfois qu'un petit nombre d'observations pour ce genre de population, la combinaison des données de plusieurs cycles est une solution que l'on peut envisager. Ainsi, elle a été utilisée par Tremblay et coll.[1] pour examiner la relation entre l'indice de masse corporelle et l'ethnicité, et par Tjepkema[2] dans une étude de l'utilisation des soins de santé par les gais, les lesbiennes et les bisexuels au Canada.

La combinaison de plusieurs cycles est possible parce qu'en général, des données sur les mêmes caractéristiques sont recueillies au cours de tous les cycles .1, et que certains renseignements identiques sont recueillis durant les cycles .2. Néanmoins, au fur et à mesure que l'ESCC a évolué, des différences sont apparues d'un cycle à l'autre, si bien que la combinaison des cycles pourrait être irréalisable, ou, si elle reste possible, pourrait affecter les résultats selon les objectifs analytiques de l'étude.

Le présent article explique les méthodes applicables pour combiner les cycles de l'ESCC et offre des lignes directrices concernant l'interprétation des résultats. Bien que l'information se rapporte spécifiquement à l'ESCC, le champ d'application de nombreuses questions est beaucoup plus vaste. Une étude de cas illustre les méthodes et

montre qu'il est possible de produire des estimations satisfaisantes d'après les données de cycles combinés.

En 2007, le programme de l'ESCC a mis en œuvre un processus de collecte continu des données dans le but de produire des fichiers annuels, ainsi que des fichiers combinés portant sur deux ans. Cette initiative a donné lieu à l'introduction de différentes « estimations de période », qui seront le sujet d'un article connexe. Le présent article est axé sur la méthodologie et les éléments à prendre en considération pour combiner les données des cycles *passés* de l'ESCC.

Une enquête évolutive

L'ESCC n'a pas été conçue comme une enquête réalisée auprès d'un échantillon avec renouvellement[3,4] construit de manière à pouvoir combiner les données recueillies au cours du temps auprès des échantillons successifs. Par conséquent, la combinaison des données ne devrait être entreprise que s'il est établi que les estimations basées sur les données d'un seul cycle ne répondent pas aux besoins analytiques et, en outre, que les résultats combinés seront pertinents et interprétables.

Depuis son lancement en 2000-2001, l'ESCC a évolué, si bien que les estimations calculées d'après les données de cycles différents ne sont pas nécessairement comparables. Pour déterminer si les cycles sont combinables, il faut tenir compte des modifications apportées au contenu du questionnaire, au champ d'observation de l'enquête, à la géographie et au mode de collecte.

Modification du contenu

Le questionnaire de l'ESCC a subi des modifications continuelles, y compris l'introduction de nouveaux modules et la suppression d'anciens. En général, si les modifications apportées au contenu sont importantes, les noms des variables changent. Néanmoins, un même nom de variable ne signifie pas nécessairement que l'on a posé la même question, de sorte que l'énoncé des questions doit être vérifié avant de combiner divers cycles. Les utilisateurs peuvent consulter la documentation sur l'ESCC, notamment les dictionnaires de données et les questionnaires, qui sont accessibles sur le site Web de Statistique Canada (enquêtes et programmes statistiques dans le module Définitions, Sources de données et Méthodes à http://www.statcan.gc.ca/concepts/index-fra.htm). Les révisions de l'énoncé des questions, de la structure des modules et des catégories de réponse peuvent signifier que la combinaison des données est inappropriée.

Modification du champ d'observation

Les populations visées par certains modules du questionnaire de l'ESCC peuvent varier d'un cycle à l'autre. L'exemple le plus évident est celui du contenu optionnel que les régions sociosanitaires ou les provinces choisissent. Par conséquent, les modules administrés aux résidents d'une région particulière durant un cycle donné peuvent être posés aux résidents d'une région entièrement différente au moment du cycle suivant.

Il se peut aussi que la population cible d'un module change. Par exemple, au cycle 1.1, les questions du module sur le comportement sexuel ont été posées aux personnes de 15 à 59 ans, mais au cycle 2.1, le groupe d'âge cible a été réduit aux 15 à 49 ans.

Modification de la géographie

Pour chaque cycle de l'ESCC, le fichier de données contient les codes et les identificateurs géographiques correspondant aux régions sociosanitaires telles qu'elles étaient au moment où les données ont été diffusées. Cependant, la délimitation des régions sociosanitaires peut changer d'un cycle à un autre. Parfois il s'agit de modifications aussi mineures que des changements de nom ou de code, mais il peut arriver que les limites des régions soient redéfinies. Le cas échéant, les fichiers doivent être mis à jour en se fondant sur une géographie commune (habituellement la plus récente) avant de pouvoir combiner les divers cycles. Plus de renseignements sur les modifications des limites sont disponibles dans la publication en ligne intitulée *Indicateurs de la santé* (section sur les régions sociosanitaires et les groupes de régions homologues, sous-section des changements apportés aux régions sociosanitaires) à http://www.statcan.gc.ca/bsolc/olc-cel/francais/olc-cel?catno=82-221-XIF&lang=fra. Si les limites des régions sociosanitaires doivent être mises à jour, les fichiers de correspondance fournissant la relation entre les aires de diffusion (AD) ou les secteurs de dénombrement (SD) et les régions sociosanitaires pour une période de référence donnée peuvent être consultés dans la publication en ligne, *Régions sociosanitaires : Limites et correspondance avec la géographie du recensement*, à http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=82-402-X&CHROPG=1&lang=fra.

Modification du mode de collecte

L'« effet de mode » est l'incidence qu'a la méthode de collecte sur la façon dont les répondants répondent à l'enquête. Les interviews de l'ESCC sont effectuées par téléphone, ainsi que sur place. L'information que les répondants fournissent peut varier selon le mode de collecte utilisé pour l'interview. Une étude de 2004[5] a révélé que plusieurs variables de l'ESCC sont sujettes à l'effet de mode, dont, sans s'y limiter, le poids et la taille, l'activité physique, les visites chez le médecin et les besoins non satisfaits de soins de santé.

Pour s'assurer que les estimations soient cohérentes, des efforts sont faits en vue de maintenir la même combinaison d'interviews par téléphone et sur place d'un cycle à l'autre. Cependant, les ajouts importants à l'enquête (achat d'unités d'échantillonnage supplémentaires) peuvent affecter l'équilibre entre les

interviews par téléphone et sur place, parce que ces interviews supplémentaires sont habituellement réalisées par téléphone. Pour le cycle 1.1, la proportion d'interviews téléphoniques était assez faible, facteur dont il devrait être tenu compte lorsque l'on envisage de combiner ce cycle à d'autres.

Combinaison d'enquêtes différentes

Pour les raisons susmentionnées, les résultats d'enquêtes sur la santé transversales *différentes* ne sont pas nécessairement comparables et, dans la plupart des situations, ne devraient pas être combinés. Par conséquent, il est conseillé de ne pas combiner la composante régionale de l'ESCC (cycles .1) avec les composantes provinciales (cycles .2 – Santé mentale (2002) et Nutrition (2004)).

Une population en évolution

La possibilité de combiner les cycles de l'ESCC découle du fait que, si des échantillons aléatoires sont tirés à partir d'une population, les échantillons cumulés peuvent être considérés comme un grand échantillon aléatoire provenant de la même population. Cependant, si la population change considérablement entre les cycles, les échantillons ne peuvent pas être traités comme s'ils provenaient de la même population. Dans le cas de l'ESCC, les échantillons des cycles successifs sont tirés à partir d'une population en évolution. Par conséquent, l'échantillon combiné n'est pas nécessairement représentatif de l'une des populations représentées par un seul cycle, mais plutôt de la population combinée.

Les différences qui se dégagent d'un cycle à l'autre peuvent être dues aux changements susmentionnés — modifications du questionnaire, du champ d'observation et du mode de collecte — ou à la variabilité d'échantillonnage. Cependant, les changements observés d'un cycle à l'autre peuvent refléter une variation réelle du paramètre étudié. Dans de telles conditions, la combinaison des cycles reste possible, mais pour interpréter les résultats, il faut comprendre l'effet des périodes que couvre l'estimation fondée sur l'échantillon combiné. Il importe aussi d'être conscient que ce genre de tendances sont obscurcies quand elles sont combinées en une estimation unique.

Méthodes de combinaison

Les méthodes de combinaison des données provenant d'enquêtes différentes peuvent être réparties en deux grandes catégories : l'approche individuelle et l'approche groupée. L'approche individuelle s'appuie sur des techniques d'estimation composite dans lesquelles les estimations sont calculées pour chaque enquête séparément, puis combinées. L'approche groupée consiste à combiner les microdonnées recueillies auprès des divers échantillons et de traiter l'ensemble de données résultant comme s'il correspondait à un échantillon tiré d'une seule population.

Approche individuelle

L'approche individuelle donne une moyenne des estimations calculées d'après les divers cycles de l'ESCC. L'avantage est que, moyennant certaines hypothèses, le résultat combiné est facile à interpréter. En outre, une moyenne peut être calculée à partir des tableaux existants, ce qui rend l'approche intéressante pour les utilisateurs des fichiers de microdonnées à grande diffusion (FMGD) et pour ceux qui s'appuient sur les tableaux d'estimations existants.

L'inconvénient de l'approche individuelle est qu'elle peut être fastidieuse. Si les estimations requises ne sont pas publiées ou ne sont pas assorties de leur variance, il faut les calculer séparément à partir de chaque enquête avant de les intégrer. Les utilisateurs des FMGD seront limités par l'information contenue dans le fichier et ceux qui s'appuient sur les tableaux devront obtenir l'accès aux microdonnées. Si de nombreuses estimations doivent être calculées, le processus prend beaucoup de temps.

Dans le cas de l'ESCC, les estimations d'un paramètre de population θ (qui peut correspondre à toute statistique, telle qu'une moyenne, un total ou un ratio) peuvent être calculées séparément pour chaque cycle, $\hat{\theta}_1$, $\hat{\theta}_2$, ... $\hat{\theta}_k$, où k est le nombre de cycles disponibles. Une simple moyenne peut alors être calculée sous la forme :

$$\hat{\theta}_c^{avg} = \frac{\sum_{i=1}^{k} \hat{\theta}_i}{k}$$

Pour pouvoir estimer assez facilement la variance, les échantillons doivent être indépendants, ce qui est le cas pour la plupart des cycles de l'ESCC. Font exception les cycles 2.1 et 2.2, où les répondants du cycle 2.1 ont été utilisés comme base de sondage pour le cycle 2.2. Par conséquent, il n'est pas facile de combiner les données des cycles 2.1 et 2.2 suivant l'approche individuelle.

Sous l'hypothèse d'indépendance entre les cycles, une estimation de la variance de la moyenne simple des trois cycles .1 peut être calculée comme il suit :

$$\hat{V}(\hat{\theta}_c^{avg}) = \hat{V}\left(\frac{\hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3}{3}\right)$$

$$= \frac{1}{9}\left[\hat{V}(\hat{\theta}_1) + \hat{V}(\hat{\theta}_2) + \hat{V}(\hat{\theta}_3)\right]$$

Il est évident que la variance estimée de la moyenne des trois cycles est approximativement égale au tiers de la variance estimée d'une estimation provenant d'un seul cycle. Les erreurs types peuvent être calculées en prenant la racine carrée de la variance et les estimations du CV peuvent être obtenues par :

$$CV(\hat{\theta}_c^{avg}) = \frac{\sqrt{\hat{V}(\hat{\theta}_c^{avg})}}{\hat{\theta}_c^{avg}}$$

Dans certains cas, il est souhaitable d'estimer une moyenne pondérée plutôt qu'une simple moyenne, en accordant

plus de poids à une estimation qu'à une autre. Si un chercheur veut estimer le même paramètre θ que celui décrit pour la moyenne simple, il peut produire des estimations distinctes $\hat{\theta}_1$, $\hat{\theta}_2$, ... $\hat{\theta}_k$ puis calculer une estimation composite ou une moyenne pondérée sous la forme :

$$\hat{\theta}_c = \sum_{i=1}^{k} \alpha_i \hat{\theta}_i$$

où $\sum_{i=1}^{k} \alpha_i = 1.$

Si chaque estimation $\hat{\theta}_i$ est une estimation sans biais de θ, alors $\hat{\theta}_c$ sera également sans biais, pour tout choix de α_i. Autrement dit, si chaque cycle donne une estimation correcte de la même statistique constante pour la même population, le résultat combiné sera une estimation correcte de la même statistique.

Suivant l'analyse que l'on veut effectuer, plusieurs choix sont possibles pour α_i. Certains d'entre eux incluent une fonction de pondération croissante, qui accorde plus de poids aux cycles les plus récents, ou une fonction de pondération basée sur les variances, qui produit une estimation plus efficace du paramètre de population (c.-à-d. ayant une variance plus faible). Pour de plus amples renseignements au sujet de ces méthodes, consulter Chu, Brick et Kalton[6], ainsi que Korn et Graubard[7].

Une fois que l'estimation composite est calculée en utilisant la valeur appropriée de i, il est possible de calculer une estimation de la variance sous forme d'une fonction des variances originales, puis d'estimer les erreurs types et les CV. Si l'on suppose que les cycles sont indépendants, la variance peut être estimée par :

$$\hat{V}(\hat{\theta}_c) = \hat{V}\left(\sum_{i=1}^{k} \alpha_i \hat{\theta}_i\right) = \sum_{i=1}^{k} \alpha_i^2 \hat{V}(\hat{\theta}_i)$$

Pour que l'approche individuelle produise une estimation sans biais d'un paramètre de population, les estimations que l'on combine doivent chacune être des estimations sans biais du même paramètre de population. Comme il est mentionné plus haut, cela pose un problème dans le cas de l'ESCC, dont le but est de mesurer les caractéristiques d'une population en évolution à différents points dans le temps. Comme l'hypothèse d'une statistique constante est douteuse, ce qui rend la moyenne pondérée difficile à interpréter, il est recommandé aux utilisateurs qui s'intéressent à l'approche individuelle d'employer la moyenne simple, qui ne requiert pas cette hypothèse ou dont le résultat est plus facile à interpréter.

L'approche groupée

L'approche groupée consiste à combiner les divers cycles de l'ESCC au niveau des microdonnées pour obtenir un ensemble de données qui peut être analysé comme s'il avait été recueilli auprès d'un seul échantillon d'une population. Cette approche est une option séduisante, parce qu'elle permet d'accroître la taille de l'échantillon et parce qu'après avoir combiné les données, il n'est pas nécessaire de retourner aux ensembles de données individuels.

Les inconvénients tiennent au fait que la manipulation des fichiers de données demande de plus grandes compétences techniques et qu'il ne s'agit pas d'une option pour les utilisateurs qui n'ont pas accès aux fichiers de microdonnées. Les utilisateurs des FMGD sont capables de calculer une estimation suivant l'approche groupée, mais ne peuvent pas calculer la variance parce que les tableaux des CV ne sont pas disponibles pour le fichier de données combinées.

Sous sa forme la plus élémentaire, un groupement des données consiste à prendre les fichiers de données individuels et les poids correspondants, puis à utiliser une simple instruction de fusion (merge) ou d'ensemble (set) dans SAS pour créer un fichier unique. Parallèlement, les fichiers de poids *bootstrap* doivent être combinés pour estimer la variance. Le fichier de données et le fichier de poids *bootstrap* résultants peuvent alors être traités comme s'il s'agissait d'un seul échantillon provenant d'une seule population. Des estimations de taux et de proportions, ainsi que des modèles statistiques, peuvent être créées à l'aide des fichiers et de tout programme statistique capable d'estimer les variances en utilisant la méthode du *bootstrap*, tel que le programme Bootvar de Statistique Canada.

L'approche décrite plus haut pourrait ne pas convenir pour estimer des totaux. Par exemple, pour estimer le nombre de cas de diabète d'après deux enquêtes indépendantes auprès d'une même population, il est impossible d'additionner les poids d'échantillon provenant des deux enquêtes pour les répondants diabétiques, car cela surestimerait le total d'un facteur deux[7]. Une option consiste à rééchelonner les poids d'échantillonnage originaux w_i en les multipliant par le facteur α_i afin de représenter la population d'intérêt, comme cela a été fait dans l'approche individuelle.

Plusieurs choix de α_i[8] sont possibles. Comme l'hypothèse selon laquelle chaque cycle de l'ESCC peut être utilisé pour estimer le même paramètre de population est douteuse, il est conseillé de rééchelonner les poids en les multipliant par un facteur constant, α_i=1/k. Si les données de deux cycles sont combinées, cela signifie que α_i=0,5; dans le cas de trois cycles, α_i=0,33. L'estimation résultante doit être interprétée comme représentant les caractéristiques de la population moyenne (ou une estimation de période), qui couvre les périodes combinées des cycles individuels. Dans ces conditions, l'hypothèse que le paramètre estimé est le même à chaque cycle n'est pas requise.

Il n'est pas toujours nécessaire d'ajuster les poids lorsque l'on regroupe les données. Si les poids sont corrigés, l'hypothèse est qu'ils le sont afin de représenter exactement une population. Le problème tient au fait que, si l'on combine des poids provenant de périodes différentes, les poids résultants ne représentent pas la population courante,

mais plutôt une population moyenne qui n'existe pas. Par conséquent, il pourrait être inapproprié de produire des totaux à l'aide d'un fichier combiné, que les poids soient ajustés ou non. Par ailleurs, les ratios, les proportions et les moyennes peuvent être des statistiques utiles si elles sont considérées comme des estimations de période. Pour ces types de statistiques, les résultats obtenus en utilisant les poids originaux ou les poids ajustés à l'aide d'un facteur commun $\alpha_i = 1/k$ seront les mêmes. Il en est également ainsi pour les paramètres de régression, les poids étant utilisés dans le modèle afin de tenir compte du plan de sondage plutôt que pour produire des estimations pour une certaine population finie.

L'une des principales applications de l'approche groupée est l'analyse complexe fondée sur des modèles de régression[1,2]. Grâce à la plus grande taille d'échantillon disponible lorsqu'on utilise des données combinées, il est possible d'étudier des modèles de régression plus détaillés. En outre, l'effet de cycle/période peut être pris en considération dans le modèle et, s'il est significatif, neutralisé. D'autres facteurs, tels que l'effet de mode peuvent également être pris en considération et neutralisés dans ce genre de modèle, ce qui permet de combiner les résultats provenant de cycles différents qui, autrement, ne seraient pas comparables.

Comparaison des approches

L'approche individuelle et l'approche groupée ne produisent pas toujours la même estimation. Par exemple, le résultat de l'approche individuelle consistant à calculer une simple moyenne de deux ratios, a/b et c/d, n'est pas égal à celui de l'approche groupée, où l'on calcule une estimation sur une période, parce que, généralement parlant

$$\left(\frac{a}{b}+\frac{c}{d}\right) \neq \frac{(a+c)}{(b+d)}$$

Par conséquent, même si les deux méthodes sont valides, le choix dépend du but de l'analyse. Dans le cas d'une estimation pour l'ensemble du Canada par exemple, certains chercheurs pourraient choisir d'étudier la moyenne des estimations provinciales, ce qui revient à attribuer le même poids à chaque province (approche individuelle), tandis que d'autres pourraient vouloir examiner l'estimation nationale (approche groupée), qui est influencée davantage par les grandes provinces que par les petites.

Dans le cas de ratios tels qu'une proportion, les deux approches produisent généralement les mêmes résultats à condition que le paramètre estimé demeure constant entre les deux occurrences, ou que la population ne change pas. Pour des statistiques comme les paramètres de régression, il est parfois préférable d'utiliser une approche groupée pour calculer les paramètres au lieu de faire la moyenne des paramètres de régression calculés pour les divers cycles.

Le projet de Durham

En 2007, la circonscription sanitaire de Durham (Ontario) a proposé de produire un rapport sur la santé des adolescents de la région en se servant des données combinées de l'ESCC. Il était prévu de prendre pour cible de l'*Adolescent Health Snapshot* le groupe des 12 à 19 ans et, dans la mesure du possible, les groupes des 12 à 14 ans et des 15 à 19 ans séparément. En se basant sur les données combinées de l'ESCC, les taux pour Durham seraient comparés aux taux provinciaux afin de révéler des différences ne se dégageant pas d'après les données d'un cycle seulement.

Les variables d'intérêt (généralement, des caractéristiques dont la prévalence est faible) étaient les suivantes :

- fumeurs quotidiens
- fumeurs quotidiens et occasionnels
- consommateurs d'alcool au moment de l'enquête
- grands buveurs
- activité sexuelle
- niveau d'activité physique
- inactivité physique
- consommation de fruits et de légumes
- utilisation d'équipement de protection (port du casque à bicyclette)
- embonpoint et obésité (indice de masse corporelle (IMC) des jeunes).

Après une première analyse afin de s'assurer que des données comparables provenant de plus d'un cycle de l'ESCC étaient disponibles, deux variables ont été éliminées :

Tableau 1
Estimations de la prévalence des fumeurs quotidiens de 12 à 19 ans, Enquête sur la santé dans les collectivités canadiennes, cycles 1.1 à 3.1, région sociosanitaire de Durham

	Cycle 1.1			Cycle 2.1			Cycle 3.1		
	Nombre dans l'échantillon	Estimation	Coefficient de variation	Nombre dans l'échantillon	Estimation	Coefficient de variation	Nombre dans l'échantillon	Estimation	Coefficient de variation
Total, groupe des 12 à 19 ans	187	61 220	...	210	66 523	...	214	70 380	...
Fumeurs quotidiens	27	7 577	22,33%	18	4 598	29,30%	16	5 110	30,26%
Proportion	...	12,38%	22,33%	...	6,91%	29,30%	...	7,26%	30,26%

... n'ayant pas lieu de figurer
Source : Enquête sur la santé dans les collectivités canadiennes, 2000-2001 (cycle 1.1); Enquête sur la santé dans les collectivités canadiennes, 2003 (cycle 2.1); Enquête sur la santé dans les collectivités canadiennes, 2005 (cycle 3.1).

- équipement de protection, à cause de modifications apportées au questionnaire entre les cycles;
- IMC, parce que la variable dérivée créée pour le cycle 3.1 n'était pas disponible pour les cycles 1.1 et 2.1.

Plusieurs autres variables possibles n'ont pas été incluses, parce qu'elles n'avaient pas été sélectionnées systématiquement comme contenu optionnel par la région de Durham. Ces variables sont les pensées suicidaires, l'insécurité alimentaire et la consommation de drogues illicites. (Un avantage secondaire du projet est qu'il a démontré la valeur de la combinaison de données de plusieurs cycles, ce qui pourrait influencer la sélection du contenu optionnel des régions dans l'avenir.)

La variable *fumeurs quotidiens* illustre le processus de combinaison des cycles. Pour toute analyse, il est recommandé de disposer d'au moins dix observations pour la caractéristique étudiée avant de calculer une estimation. Même en combinant les données, l'analyse du groupe des 12 à 14 ans a été impossible, à cause de la taille limitée de l'échantillon et du petit nombre de répondants qui fumaient quotidiennement. Cependant, il a été possible d'examiner la consommation quotidienne de cigarettes chez les jeunes de 15 à 19 ans de la région de Durham.

L'analyse préliminaire des données pour le groupe complet des 12 à 19 ans a consisté à calculer les estimations pour chaque cycle individuellement. Les résultats ont indiqué clairement que pour le groupe d'âge complet, la combinaison des cycles n'était pas nécessaire : pour chaque cycle, les estimations de la prévalence des fumeurs quotidiens étaient publiables, les coefficients de variation étant inférieurs au seuil d'exclusion recommandé de 33 % (tableau 1). L'analyse a également montré que la proportion de fumeurs quotidiens chez les 12 à 19 ans avait baissé fortement, pour passer d'un peu plus de 12 % au cycle 1.1 à environ 7 % aux cycles 2.1 et 3.1. Par conséquent, il aurait été incorrect de conclure que les taux étaient les mêmes d'un cycle à l'autre, ce qui rendait donc inappropriées certaines méthodes de combinaison des cycles décrites plus haut. En outre, la baisse du taux d'usage du tabac donnait à penser qu'il ne convenait peut-être pas de combiner les données du cycle 1.1 à celles des autres cycles. Si cette chute du taux reflète une initiative stratégique importante, il serait préférable d'analyser les données combinées uniquement pour les périodes durant lesquelles la politique était en vigueur (cycles 2.1 et 3.1).

L'approche individuelle consistant à calculer une moyenne simple et l'approche groupée consistant à calculer une estimation de période ont été utilisées l'une et l'autre pour combiner les données des trois cycles. Les données combinées masquent les changements de comportement, notamment la réduction importante de l'usage du tabac chez les adolescents. En outre, l'estimation résultante prête à confusion, puisqu'elle diffère des taux publiés les plus récents. Pour ces raisons, il faut interpréter les estimations comme étant des moyennes de période plutôt que reflétant les taux actuels d'usage du tabac.

Dans le cas de l'*approche individuelle*, on a calculé la moyenne des estimations du pourcentage de fumeurs quotidiens :

(12,37 % + 6,91 % + 7,26 %) / 3 = 8,67 %.

Pour estimer la variance, on a estimé la variance pour chaque cycle. Pour le cycle 1.1, la variance estimée a été calculée comme il suit :

Variance estimée = $(CV*Estimation)^2$ = $(0,2233*0,1238)^2$ = 0,0008.

Des estimations comparables ont été calculées pour les cycles 2.1 et 3.1, soit 0,0004 et 0,0005, respectivement. Ces variances estimées ont ensuite été utilisées pour estimer la variance de l'estimation combinée, soit

Variance combinée estimée = (0,0008 + 0,0004 + 0,0005) / 9 = 0,0002.

Le CV pour l'estimation combinée a été calculé par

CV combiné = racine carrée (0,002) / 0,0867 = 16,3 %,

ce qui représente une amélioration par rapport aux CV calculés pour un cycle seulement et est acceptable pour la diffusion, conformément aux lignes directrices concernant la publication.

Dans le cas de l'*approche groupée*, on a calculé une estimation de la période :

(7 577 + 4 598 + 5 110) / (61 220 + 66 523 + 70 380) = 17 285 / 198 123 = 8,72 %.

Le faible écart entre la moyenne simple et l'estimation de période est dû principalement à des variations de la taille de la population et du taux d'usage du tabac.

Pour l'approche groupée, les poids auraient pu être ajustés en divisant les poids originaux par trois, mais le résultat aurait été le même :

5 761 / 66 041 = 8,72 %.

Par contre, dans le cas des totaux, la population estimée était de 198 123 en utilisant les poids non ajustés, ce qui correspond approximativement à trois fois l'estimation pour chaque cycle. L'estimation groupée en utilisant les poids ajustés était 66 041, soit la moyenne des chiffres de population pour chaque cycle.

Pour estimer les variances dans le cas de l'approche groupée, on a utilisé Bootvar pour calculer les estimations par la méthode du *bootstrap*. L'estimation de la variance de l'estimation groupée était 0,0002, avec un CV correspondant de 15,3 %. Comme il est montré dans le cas de l'approche individuelle, il s'agit d'une amélioration par rapport aux estimations obtenues en traitant chaque cycle individuellement.

Enfin, nous nous attendions à ce qu'une comparaison des taux groupés pour la région de Durham au taux provincial révèle des écarts statistiquement significatifs, à cause de la plus grande précision due à l'accroissement de la taille de l'échantillon. En général, cela n'a pas

été le cas. Les écarts entre les taux pour l'Ontario et pour Durham étaient si faibles qu'il n'était pas possible de les déceler, même avec les tailles d'échantillon plus grandes.

Conclusion

La combinaison des cycles de l'ESCC produit des échantillons de plus grande taille pour l'analyse et les estimations résultantes sont de meilleure qualité que celles obtenues en utilisant les données d'un seul cycle. Néanmoins, on ne peut supposer que les estimations résultantes représentent la même population ou que les caractéristiques de la population sont les mêmes que celles qui se dégageraient de l'analyse des données d'un cycle seulement, même si la même question a été posée d'un cycle à l'autre. Au fil du temps, les personnes qui constituent la population et leurs caractéristiques évoluent. Les estimations basées sur les cycles combinés décrivent une population « artificielle » constituée de populations différentes étudiées à des moments différents. Par conséquent, les chercheurs doivent déterminer quelles sont les implications pour leurs analyses avant de décider de combiner les données de plusieurs cycles.

Références

1. M. Tremblay, C. Pérez, C. Ardern *et al.*, « Obésité, embonpoint et origine ethnique », *Rapports sur la santé*, 16(4), 2005, p. 25-37 (Statistique Canada, n° 82-003 au catalogue).
2. M. Tjepkema, « Utilisation des services de santé par les gais, les lesbiennes et les bisexuels au Canada », *Rapports sur la santé*, 19(1), 2008, p. 57-70 (Statistique Canada, n° 82-003 au catalogue).
3. L. Kish, « Le cumul ou la combinaison d'enquêtes démographiques », *Techniques d'enquête*, 25(2), 1999, p. 147-158 (Statistique Canada, n° 12-001 au catalogue).
4. C.H. Alexander, « Les échantillons successifs de Leslie Kish et l'American Community Survey », *Techniques d'enquête*, 28(1), 2002, p. 35-41 (Statistique Canada, n° 12-001 au catalogue).
5. M. St-Pierre et Y. Béland, « Mode effects in the Canadian Community Health Survey: a comparison of CAPI and CATI », *Proceedings of the American Statistical Association Meeting, Survey Research Methods*, Toronto, American Statistical Association, 2004.
6. A. Chu, J.M. Brick et G. Kalton, « Weights for combining surveys across time or space », *Bulletin of the International Statistical Institute: 52nd Session, Contributed Papers, Book 2*, 1999, p. 103-104.
7. E.L. Korn et B.I. Graubard, « *Analysis of Health Surveys* », New York, Wiley, 1999.
8. E.M. Friedman, D. Jang et V.T. Williams, « Combined Estimates from Four Quarterly Survey Data Sets », *Proceedings from the Joint Statistical Meetings – Section on Survey Research Methods*, 2002, p. 1064-1069.